ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 13 (2023) No. 6, pp. 230-243
DOI:10.33168/JSMS.2023.0614

A Machine Learning Approach for Model Selection of Social Aid Beneficiaries

Mahfudh Ahmad*, Harjanto Prabowo, HL Hendric Spits Warnars, Ford Lumban Gaol

Computer Science Department, BINUS Graduate Program, Bina Nusantara University, Jakarta, 11480, Indonesia

mahfudh@binus.ac.id (corresponding author)

Abstract. Identifying and providing well-targeted social assistance to potential beneficiaries for poverty alleviation is important for achieving the Sustainable Development Goals (SDGs) targets. Machine learning algorithms can be a solution in achieving SDGs. This paper aims to model poverty measurement based on integrated data using machine learning, supervised methods can classify family and household information with high precision and show features or indicators of various dimensions of poverty that are comparable to learning methods. We used the Proxy Means Test (PMT) for labeling the data obtained from national targeting database, along with Linear Discriminant Analysis (LDA), Light Gradient Boosting (LightGBM), Extra Trees (ET), Random Forest (RF), and Decision Tree (DT). Based on the experiment, we obtained optimal results for the five methods based on the output of the F1 score, accuracy, precision, AUC indicators, and the highest value of the Kappa coefficient. Then, we looked at how hyperparameter tuning affected various learning algorithms, and the LDA proved to be the most effective approach, with an F1 value of 0.891 percent and Kappa score of 0.8450 percent. To our knowledge, there is no research that discusses targeting data for social aid using machine learning and features reduction on the variables used yet. This paper contains research to fill this gap, and the method is accurate, affordable, and brings poverty identification closer to real-time and at a lower cost.

Keywords: machine learning, supervised classification, supervised classification, poverty measurement.

1. Introduction

Poverty is a global phenomenon that affects people or households to varying degrees everywhere in the world (Gallardo, 2018). Both our understanding of measurement approaches and poverty has considerably improved the targeting efforts of intervention types during the last decades. The governments and social organizations adopted several initiatives aimed at assisting the disadvantaged, particularly households and families. Some families are falling near the poverty line. These households typically do not get any assistance, as their incomes are a little bit above the poverty line. This set of households should be the focus of efforts to keep them out of poverty (Batana et al., 2013).

Poverty alleviation has many dimensions that must be carried out by the government, companies through Corporate Social Responsibility (CSR), and the community, but accurate data and the ability to provide an overview of poverty in a target area is an important factor (Alam, 2017). With the availability of good and accurate data, the government can use the data for planning and budgeting for poverty reduction programs and provide appropriate programs both by locus and target households/individuals. Some developing countries, due to budget constraints, require data that classify the socioeconomic status of the households within the region, i.e., poor, near-poor, poor, and very poor. When the government aims to serve the deprived one, then there should be a way to identify poor households as beneficiaries or potential beneficiaries of poverty reduction programs. Further, the government can map regional priorities and the number of targets for program intervention based on the available budget and the classified data.

During 2015 in Indonesia, the National Team for the Acceleration of Poverty Reduction (TNP2K) developed a targeted database to maintain all records on poor, near poor, and vulnerable households. To obtain a ranking of socio-economic status within the household, the TNP2K modeling team built a targeting database containing data on residents with the poor, vulnerable and poor categories and residents labeled as potential beneficiaries, TNP2K used the Proxy Means Testing (PMT) model to predict per capita expenditure and obtain welfare status in a household by using indicators in the data collection instrument as well as additional other indicators, such as the geographic accessibility index (IKG in Indonesian) and Social Economic Survey (SUSENAS) as the weight and refinement of the PMT model (Bah et al., 2014).

This study conducted a comprehensive comparative study and found the best model for the classification method to predict the welfare status of households using existing datasets, more specifically in:

1. Understanding theory of classification technique and the algorithms.
2. Reviewing some well-known experiments of the recently published journal for machine learning classification for poverty measurements.
3. Setting up a model using the classification method and enhance the score.
4. Finding and comparing accuracy, the error value of the metric, F1 Score, and execution times.

Therefore, this paper aimed to provide information and methods on the classification of poverty in Indonesia by predicting socioeconomic status with machine learning.

Inaccurate and out-of-date data is an obstacle in some countries to provide social protection/poverty alleviation benefits to the community. When the data is available, it is often difficult to predict socio-economic status/poverty so mistargeted occurred. The target household classification model can provide information quickly in predicting the socio-economic status of households. Therefore, it can assist the government in planning social assistance programs and allocating the right amount of budget to the beneficiaries.

In previous research, Han and Kamber defined classification as the process of identifying a model or function that could characterize and discriminate classes of information or ideas, with the goal of using the model to forecast an unidentified class of a seen object. Conditioning machines to learn without being explicitly programmed is the goal of the artificial intelligence (AI) technique known as machine learning (Edgar & Manz, 2017; Han & Kamber, 2012). Learning has a crucial role in what makes us human. We must create machines that can learn on their own, from their prior experiences, if we are to create AI that can do tasks with human-like intelligence. A machine learning method that can be used in classification includes Linear Discriminant Analysis (LDA), Light Gradient Boosting (LightGBM), Extra Trees (et), Random Forest (rf), K Neighbors (knn), and others (Davies, 2018).

Researchers of a variety of disciplines and areas have extensively discussed classification approaches in machine learning. This innovation gives disclosures and benefits of unused methods in information investigation and forecast. New ways for data analysis and prediction have been discovered thanks to this technology. In many fields, classification has been used as a technique of reference since it provides information on accuracy and precision, such as stock market (Nti et al., 2020; Ravikumar & Saraf, 2020), customer profiling (Noori, 2021; Palaniappan et al., 2017), image recognition, disease diagnosis (Mohammed & Al-Tuwaijari, 2021; Nilashi et al., 2020; Schaefer et al., 2020) and some machine learning method have been experimented in the social-economic domain and increasingly being popular, such as random forest (Kambuya, 2020; Otok & Seftiana, 2014; Zixi, 2021), satellite and/or phone and social analysis (Pokhriyal et al., 2020; Wang et al., 2021).

2. Method

2.1. Approach

In this work, the topic of predicting poverty was addressed using well-known machine learning methods. Before creating new algorithms for this problem, the performance of existing algorithms should be evaluated, as this research was the first of its kind in Indonesia.

The study started by analyzing and comprehending the dataset at hand, after which it discussed the issues rose by the data and offered solutions. Following data processing, 10 machine-learning algorithms were fed the data. The findings were given and discussed at the end. A full illustration of the study's flow is shown in Figure 1. The three primary elements of the strategy will be presented in further detail in the following sections.

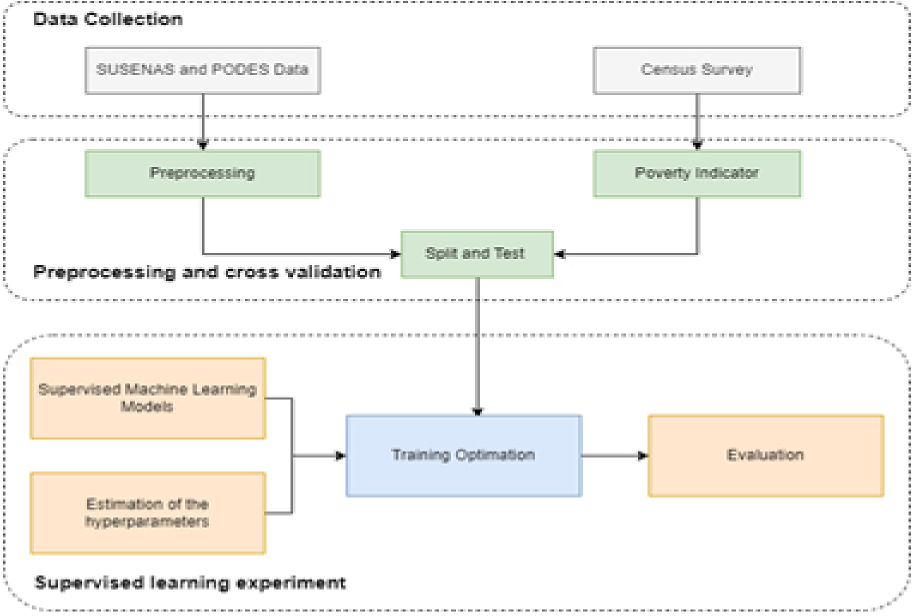

Data Collection
SUSENAS and PODES Data
Census Survey
Preprocessing
Poverty Indicator
Split and Test
Preprocessing and cross validation
Supervised Machine Learning Models
Estimation of the hyperparameters
Training Optimation
Evaluation
Supervised learning experiment

Fig.1: Illustration of the technique used

2.2 Dataset

The Unified Data became a reference for every poverty reduction program in Indonesia through data collection in 2011 and updated in 2015 by Statistics Indonesia (BPS) and managed by TNP2K. Based on published guidelines (TNP2K, 2018a), there were seven groups of variable classifications as a measure of poverty in households: expenditure, demographics, education, employment, housing facilities, asset ownership, and geographic accessibility index. Based on the results of the TNP2K evaluation (TNP2K, 2018b), Unified Database has been used by all districts and provinces in Indonesia as baseline data to achieve SDG targets in terms of reducing poverty.

Table 1. Variables description for model

Variables	**Features**	
	Type	***Description***
h_aset_car	Numerical	# Owned car
h_aset_fridge	Numerical	# Owned refrigerator
h_aset_motorcycle	Numerical	# Owned motorcycle
h_aset_perahu	Numerical	# Owned boat
h_aset_phone	Categorical	Owned phone
h_cookingfuel	Numerical	Main cooking source
h_dwater	Numerical	Drinking water source
h_hhsize	Numerical	Household size
h_lighting	Numerical	Main electricity source
h_lpcfloor	Numerical	Log of Unit Square Meter per Person
h_nage	Numerical	Age
h_ngrad	Numerical	# individu completed school and college
h_nstudents	Numerical	# individu in school
h_pwater	Numerical	Access to drinking water
h_sec1_stat	Numerical	# working members in the household
h_septic	Numerical	Disposal site
h_tfloor	Numerical	Floor size
h_toiltype	Numerical	Toilet tyoe
h_troof	Numerical	Roof type
h_twall	Numerical	Wall type
h_aset_bicycle	Numerical	# Owned bicyle
h_cookingfuel	Numerical	Cooking fuel
h_elderly	Numerical	# elders in the household
h_child	Numerical	# children in the household
h_adult	Numerical	# adults in the household
h_educ	Categorical	Education of household head
h_married	Categorical	Marital status head household
h_house	Numerical	House status ownership
Urban_rural	Categorical	Type village
h_gender	Categorical	Gender
ikg	Numerical	Geographic accessibility index
poor_status	Categorical	Is the household poor?

Figure 2 shows the distribution of each decile, where the 1st decile was 1,451, the 2nd decile was 2,122, the 3rd decile was 2,532 and the 4th decile was 1,422 households. After exploration and cleaning dataset like filling null values, removing object data unsuitable for the model, transforming categorical and numerical variables, the next step was removing the multicollinearity in the variables using the correlation matrix in Python. The point was to see if there was still a close relationship between variables as multicollinearity had a negative impact on the modeling.

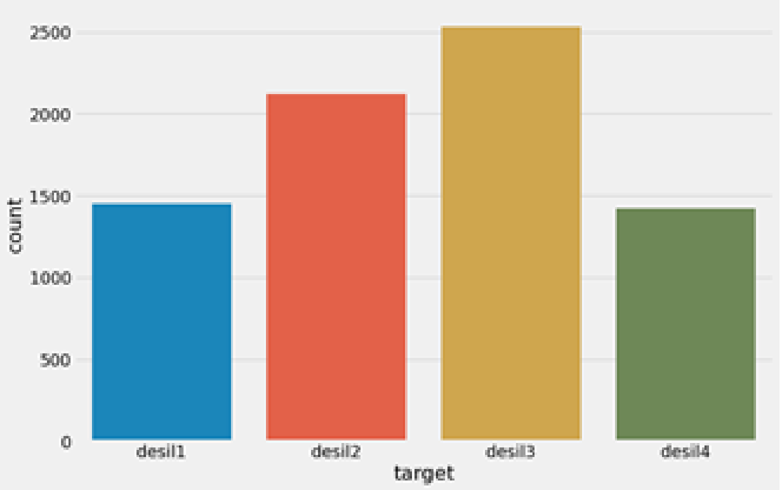


Fig. 2: Distribution of Data

2.3. Model selections

2.3.1. Linear Discriminant Analysis (LDA)

A statistical theory-based technique that has found widespread application in both machine learning and statistics is linear discriminant analysis (LDA), often known as Fisher's linear discriminant, used in terms of data processing, and image processing. LDA was first applied to the face recognition process by Etemad and Chellappa (Etemad & Chellappa, 1997). ANOVA and LDA are closely related (analysis of variance), which requires the assumption of the same variance. The identical variance-covariance matrix (between input variables) of the classes must be assumed for LDA to work. For the categorization phase of the analysis, this assumption is crucial. Fisher's theory of situations of a class is directly extended by LDA, which computes these situations using matrix algebraic techniques (such as eigendecomposition) (Fisher, 1936; McLachlan, 2005).

Although the most popular method for feature extraction is Principal Component Analysis (PCA) (Martinez & Kak, 2001), it has a weakness, in which the separation between classes is less than optimum. Therefore, the LDA method was created to overcome the shortcomings of PCA. LDA attempts to obtain combinations of features in a linear way by separating multiple objects from a class and trying to keep as much information while reducing the number of dimensions (i.e., variables) in the data set. With the LDA technique, the best combination of features linear with the separation of objects in different classes (Izenman, 2013) can be found. For example, prediction classification of individuals in psychology (Boedeker & Kearns, 2019), or obtaining information and identify whether diseases such as diabetes or hypertension are still the main measures that reduce life span (Lee et al., 2014), or a physician can analyze and identify whether the patient is in a high or low-risk stroke.

2.3.2. Random Forest (RF)

One of the supervised classification algorithms is the Random Forest (RF). As the name implies, it entails generating a forest and randomizes it. Therefore, the more trees it has in the forest, the more precise the results. However, it needs to be understood that creating a forest is different from making decisions with the information gain or gain index approach. This approach comprises of a number of tree-structured classifiers {h (x, Θk), k=1,...} where the {Θk} is an independently distributed random vector and each tree has a unit vote for the most popular class in input x. The following list contains the RFs algorithm's steps.

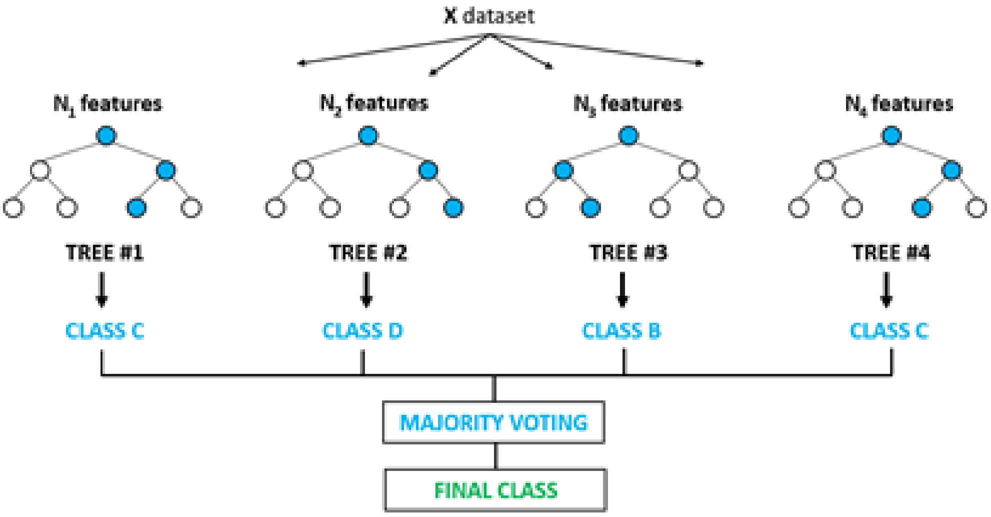


Fig.3: Random Forest

Compared to other classification algorithms, it is thought to operate quickly, better, and with a high degree of accuracy (Thoplan, 2014). Breiman was the first to formally introduce Random Forest, a mixture of models for improving and discovering categorization accuracy, after the bagging approach in 2001. Due to its enormous number of trees and the strong rule of large numbers, RF's classification model is able to avoid overfitting compared to other classification models (Breiman, 2001) , Bau presented RF could predict breast cancer as the early detection and could prevent death or receiving late treatment (Bau et al., 2022).

2.3.3 Boosting Algorithms

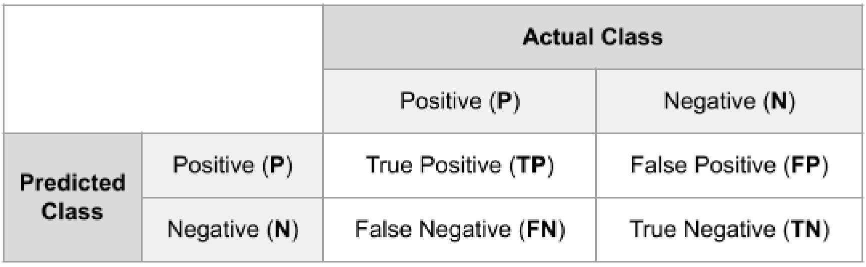
Boosting is a meta-algorithm in machine learning for supervised learning. There are various types of supervised classifiers, such as Naive Bayes Classifier, generalized linear models, linear discriminant analysis (LDA), stochastic gradient descent, and quadratic discriminant analysis (QDA), support vector machines (SVM), decision trees of linear support vector classifiers (Linear SVC), nearest neighbors, neural network models, and ensemble methods.

This predictive model's main objective is to boost overall accuracy. Two methods can be used to accomplish this. One is by employing boosting algorithms. Another one is by using feature engineering. The training observations that lead to misclassification are the focus of the boosting algorithm. There are five commonly utilized enhancement techniques, including gradient enhancement, AdaBoost, CatBoost, LightGBM, and XGBoost (Friedman, 2002; Ke et al., n.d.).

2.4 Model selections

To determine which way of partitioning the data set and categorization could create a level of accuracy, evaluation was performed. The Confusion Matrix was taken into consideration when making assessments in this study. The Confusion Matrix is a method for evaluating how well classifiers can identify or predict different types of input. Confusion Matrix is a table measuring m×m with m=number of classes (Sammut & Webb, 2010). Each class's actual label filled the column section, while the projected class label filled the row section, as shown in table 2.

Table 2. Confusion Matrix



		Actual Class	
		Positive (**P**)	Negative (**N**)
Predicted Class	Positive (**P**)	True Positive (**TP**)	False Positive (**FP**)
	Negative (**N**)	False Negative (**FN**)	True Negative (**TN**)

The percentage of the right frequency classified with the entire sample, or accuracy of classification, was typically used as a measure of accuracy. Hence, we could examine the sensitivity as well as the accuracy. The percentage of the target class that was accurately predicted was known as sensitivity +. Specificity was the percentage of classes that were correctly predicted but unimportant or undesired. When accuracy was high, but sensitivity and specificity were poor, then the categorization was unreliable.

$$\text{Accuracy} = (TP + FN)/(TP + FP + FN + TN) \tag{1}$$

$$\text{Precision} = TP/(TP + FP) \tag{2}$$

$$\text{Recall} = TP/(TP + FN) \tag{3}$$

$$\text{F-measure} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \tag{4}$$

3. Results and Discussion

The first evaluation we did was running the classification model on all data by default without paying attention to the imbalance data and not eliminating collinearity between variables as shown in Figure 4.

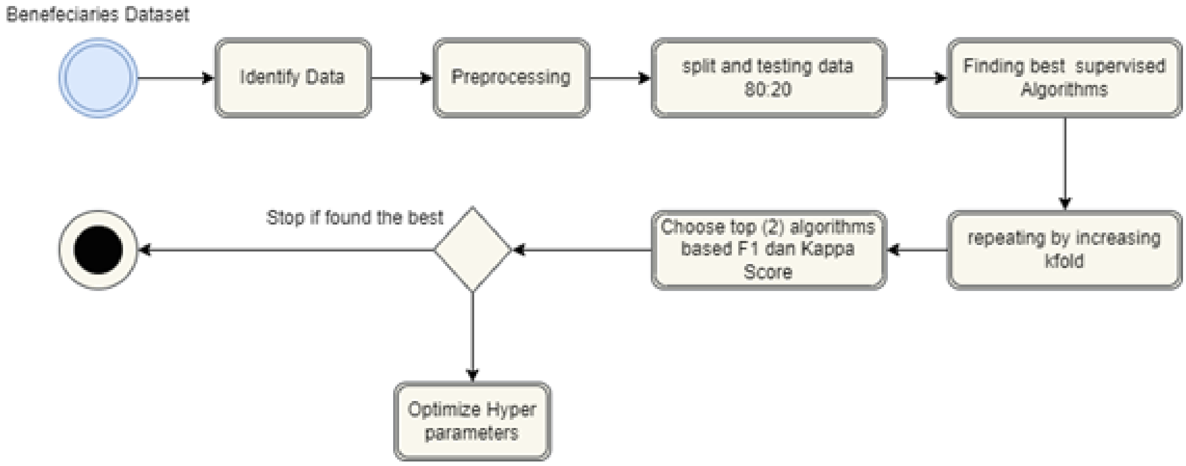

Benefeciaries Dataset
Identify Data
Preprocessing
split and testing data 80:20
Finding best supervised Algorithms
repeating by increasing kfold
Choose top (2) algorithms based F1 dan Kappa Score
Stop if found the best
Optimize Hyper parameters

Fig.4: Activity Diagram (UML) for supervised learning

By utilizing the feature selection module in Python, we found that of the 182 variables in this dataset, only 53 features could be used to get 95% accuracy. To see which features in dataset were important and impacted on the target/welfare status in the household, we used the existing SHAP module in

Python programming to optimize machine learning models and speeding up the learning process and socio-economic calculations in households.

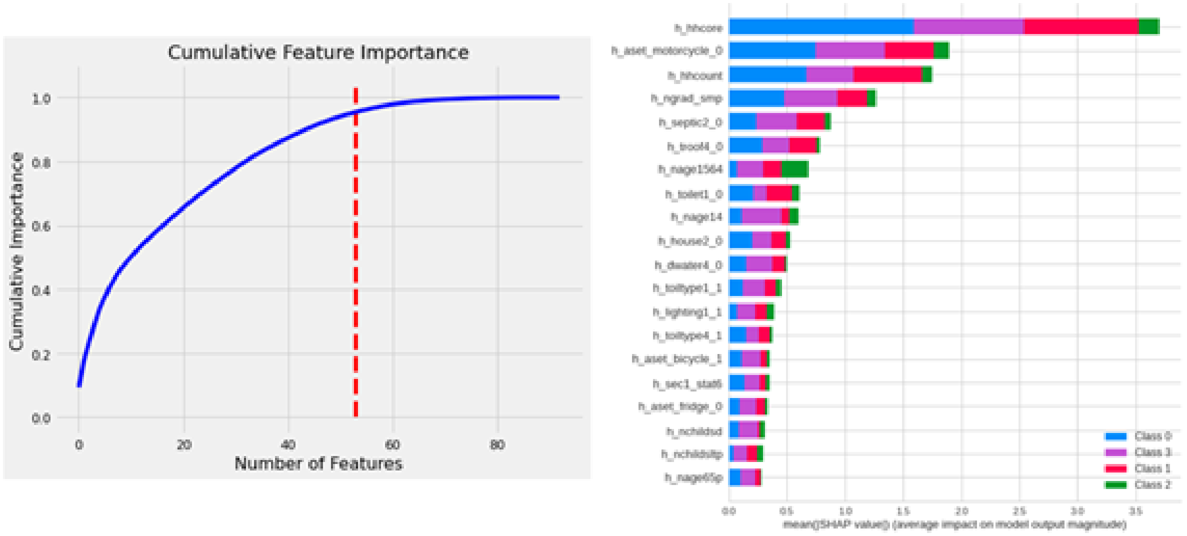


Fig.5: Number of recommended features and most impact variables

If we take a closer look to Figure 5, number of household members, ownership of motorized vehicles, number of children attending school, type of septic tank, age of children attending school, number of elderly livings in the house, home ownership, type of toilet, type of lighting, type of work, refrigerator ownership, children under five who live in the house, were important variables in determining the welfare status of the household.

3.1. Finding Best Models for Poverty Classification

In this machine learning experiment, we got the results from several supervised algorithms, where LDA and LightGBM obtained the highest accuracy among the ten methods. Therefore, we would focus on using these two to indicate the ranking of welfare status.

Table 3. Comparative of machine learning models

	Model	F1	Kappa
lda	Linear Discriminant Analysis	0.8916	0.8450
LightGBM	Light Gradient Boosting Machine	0.8044	0.7221
xgboost	Extreme Gradient Boosting	0.8046	0.7217
gbc	Gradient Boosting Classifier	0.7354	0.6263
rf	Random Forest Classifier	0.7073	0.5848
et	Extra Trees Classifier	0.7066	0.5841
dt	Decision Tree Classifier	0.6837	0.5502
ada	Ada Boost Classifier	0.6866	0.5561
ridge	Ridge Classifier	0.6195	0.5016

After cleaning the data and getting the features that had an impact on measuring poverty, we looked for the best classification model to make predictions for households. The search for the best model then would focus on the supervised classification model and choosing the top two scores based on the F1 score and Kappa score to optimize features and hyperparameters.

Based on Table 3, using Algorithms of Linear Discriminant Analysis (LDA), Light Gradient Boosting (LightGBM), Decision Tree (DT), Extra Trees (ET), Random Forest (rf), the results shown that LDA had the highest accuracy value compared to eight other models, as well as the F1 score, and the highest level of reliability from the Kappa score.

$$\kappa = 1 - \frac{\sum_{i,j} \text{weighted}_{i,j} C_{i,j}}{\sum_{i,j} \text{weighted}_{i,j} E_{i,j}} \tag{5}$$

A classification approach known as logistic regression has historically been used solely for issues with two classes. The preferred linear classification method was linear discriminant analysis if there were more than two classes. In classification, K classes—1, 2, 3,..., K – and an input vector X were taken into account.

$$\delta(X) = \frac{\mu_i X}{\sigma^2} - \frac{1}{2}\frac{{\mu_i}^2}{\sigma^2} + \log\big(P(Y = i)\big) \tag{6}$$

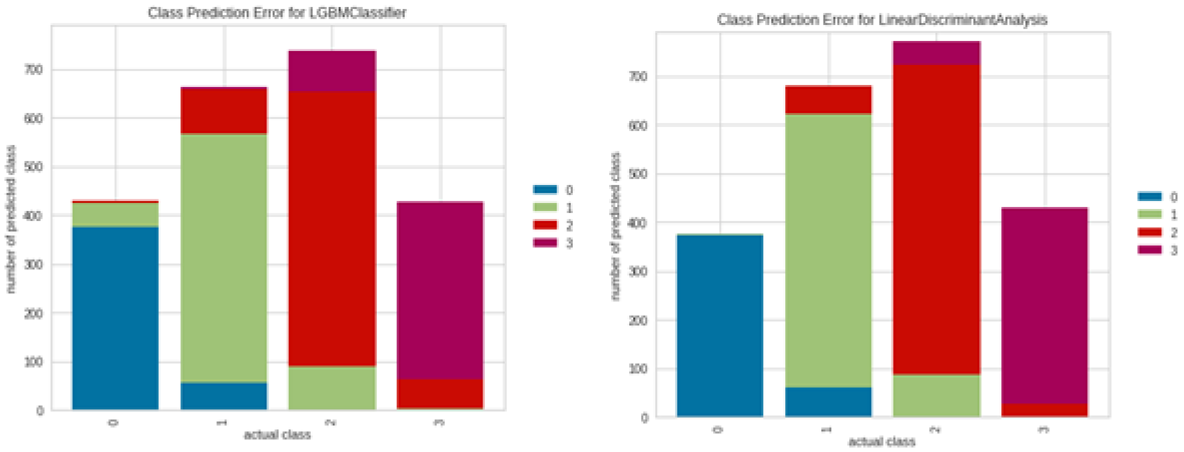
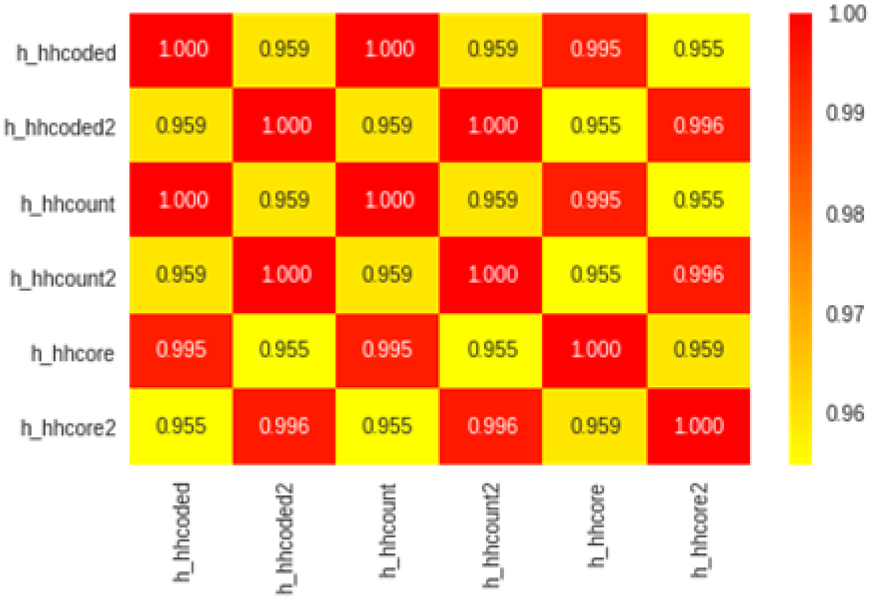


Fig.6: Prediction Error LightGBM versus LDA

From the prediction error in figures 6, we concluded that LDA was more accurate in predicting a household belonging to the category of poor/not poor. The labels of 0,1,2,3 in figure 6 were categorized as chronic poverty, moderate poverty, vulnerable, and non-vulnerable.

3.2. Removing Multicollinearities

As shown in figure 6, the output of error prediction shown that the model was more focused on finding the median of the target. As the predictions were piled up in the middle between deciles 2 and 3, we tried to reconstruct the variables by looking at the correlation between variables. There were variables that had high correlation with each other (was greater than .90) and would have an impact on the performance of the model, which indicated multicollinearities. Therefore, we then eliminated the high correlation variables.



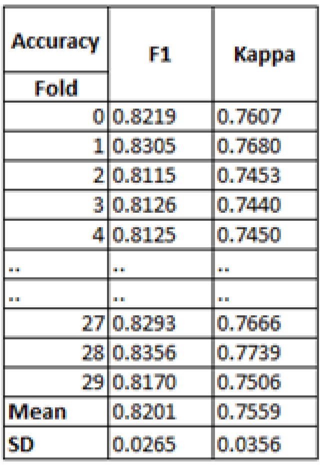
	h_hhcoded	h_hhcoded2	h_hhcount	h_hhcount2	h_hhcore	h_hhcore2
h_hhcoded	1.000	0.959	1.000	0.959	0.995	0.955
h_hhcoded2	0.959	1.000	0.959	1.000	0.955	0.996
h_hhcount	1.000	0.959	1.000	0.959	0.995	0.955
h_hhcount2	0.959	1.000	0.959	1.000	0.955	0.996
h_hhcore	0.995	0.955	0.995	0.955	1.000	0.959
h_hhcore2	0.955	0.996	0.955	0.996	0.959	1.000

Fig.7: High correlation between variables

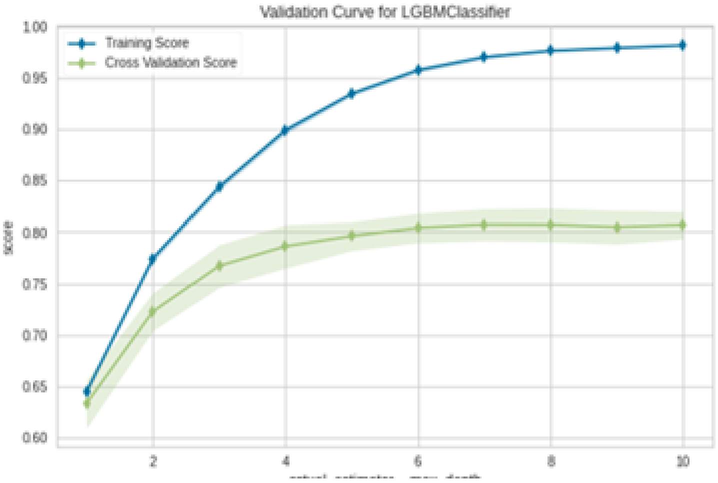
3.3. Tuning LightGBM

The results of LightGBM tuning were almost close to LDA, especially seen from the Kappa and F1 scores. These indicated that LightGBM could improve accuracy if the k-fold was increased to 30-fold. In the experiment with table 4, it was found that the LightGBM model could be increased with accuracy = 0.8201 and the Kappa score became higher (0.7559), from 0.7221. However, the maximum score obtained was 0.8219 with the highest Kappa score 0.7680, which still not exceed the LDA score.

Table 4. Fine tuning hyperparameter for LightGBM



Accuracy	F1	Kappa
Fold		
0	0.8219	0.7607
1	0.8305	0.7680
2	0.8115	0.7453
3	0.8126	0.7440
4	0.8125	0.7450
..	..	..
..	..	..
27	0.8293	0.7666
28	0.8356	0.7739
29	0.8170	0.7506
Mean	0.8201	0.7559
SD	0.0265	0.0356



The variables that we optimized, we turned into reference variables in the previously developed model, and then we ran through the same model again. In getting the best results, we used k-fold validation starting from 5,10,15,20,25, and 30. With this cross-validation, we tested the model as many as (k) times using different split training data. 10-Fold validation was commonly used in machine learning to be considered to get the best and sufficient validation and an effective method for model testing. But if the model was suboptimal, the k-fold value could be increased to 25 or 30 folds. While evaluating the model's performance, we employed the F1 Macro measure and the Kappa value to boost our level of confidence in its validity.

The model was then finalized, saved into a pickle file, and the unseen dataset was predicted. The result showed an accuracy score of 0.82 and a Kappa score of 0.75 percent.

3.4 Model Evaluation

Based on the results of comparison and model testing and model validation, LDA seemed to be the most effective model. Although the performance of hyperparameters helped improve the performance of machine learning models, we had insufficient time to test for every potential combination of settings for each model. We found that the confidence level of each class was relatively high. As can be seen in Figure 8, this model could perfectly capture socioeconomic information on decile 1.

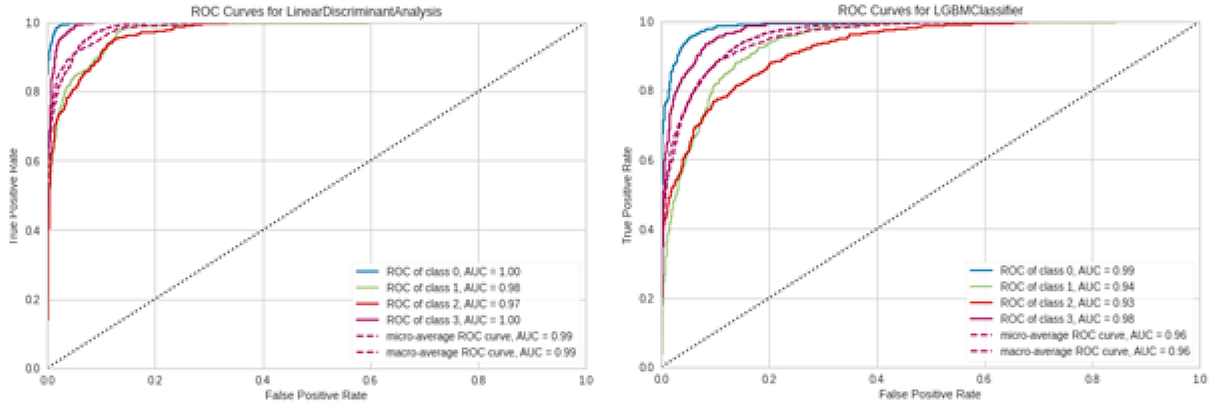


Fig.8:Receiver Operating Characteristic based on LDA and LightGBM

To determine that the data collection and categorization of the model could create a level of accuracy, the confusion matrix was considered when making the assessment. The confusion matrix in Figure 9 below provided information that LDA could see more features from decile-1 household than LightGBM, even for deciles-above-1, LDA looked better. Based on the results of comparison and model testing and model validation, LDA seemed to be the most effective model.

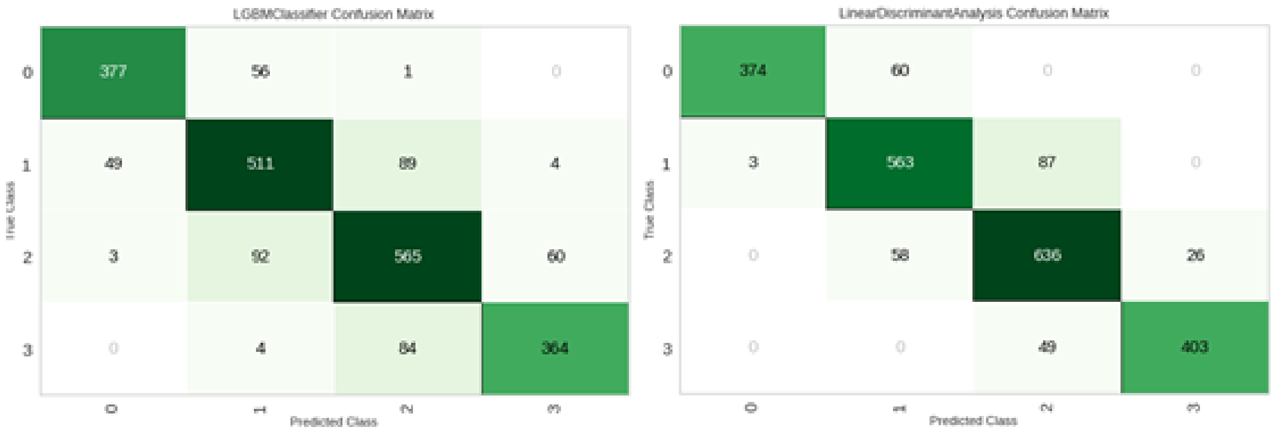
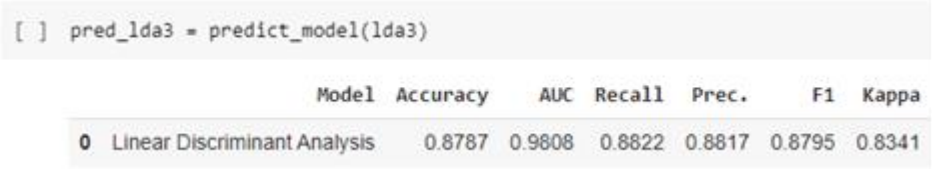


Fig.9: Confusion Matrix LightGBM versus LDA

3.5. Demonstration

The results that we observed were based on k-fold cross validation on dataset (0.8916). To see predictions and model performance on test/hold-out datasets, the prediction model function was used. The predict model function was also used to predict the latest survey dataset. At this point onward, we would use the same dataset for training as a proxy, for the new dataset did not have welfare state yet. In practice, the predict model function would be used iteratively, each time with a new data set. At this final stage, we would take a test dataset that had been prepared to make predictions on the model that had been selected.



```
[ ]  pred_lda3 = predict_model(lda3)

                          Model  Accuracy     AUC  Recall   Prec.      F1   Kappa
0  Linear Discriminant Analysis    0.8787  0.9808  0.8822  0.8817  0.8795  0.8341
```

Fig.10: Demonstration using new dataset

Table 5. Tuning Hyperparameters

Models	Features	
	Best Score	***Best Parameters***
LDA	0.8916	{'imputer': Simple_Imputer, 'dummy', Dummify, 'clean_names', Clean_Colum_Names(), solver='svd', tol=0.0001}
LightGBM	0.8201	LGBMClassifier(boosting_type='gbdt' ,colsample_bytree=1.0, importance_type='split', learning_rate=0.1, max_depth=-1, min_child_samples=20, min_child_weight=0.001, n_estimators=100, num_leaves=31, random_state=771, silent=True, subsample=1.0, subsample_for_bin=200000)

4. Conclusion

In this study, we implemented a complete data science solution to a real-world problem step-by-step. Machine learning is essentially just a set of easy-to-follow procedures that together produce an often incredibly potent final product. Although our final model performed well, there was still space to increase its accuracy, but we might not have sufficient time to attain overall great metrics. Limitation of this study is that the validity of supervised learning was highly dependent on genuine answers from respondents and input from enumerators, but using external datasets, such as electricity usage, data usage and communication with mobile phones, could minimize the risks. Although the authors only used supervised learning to rank the welfare status of households, there are still some other techniques that might prove useful.

References

Alam, K. (2017). Poverty reduction through enabling factors. *In WJSTSD*, *14*(4), 310–321. https://doi.org/10.1108/WJSTSD-07-2016-0049

Bah, A., Bazzi, S., Sumarto, S., & Tobias, J. (2014, November 12). *Finding the Poor vs. Measuring Their Poverty: Exploring the Drivers of Targeting Effectiveness in Indonesia.* [MPRA Paper]. https://mpra.ub.uni-muenchen.de/59759/

Batana, Y., Bussolo, M., & Cockburn, J. (2013). Global extreme poverty rates for children, adults and the elderly. *In Economics Letters*, *120*(3), 405–407. https://doi.org/10.1016/j.econlet.2013.05.006

Bau, Y. T., Sasidaran, T., & Goh, C. L. (2022). Improving Machine Learning Algorithms for Breast Cancer Prediction. Journal of System and Management Sciences, 12(4), Article 4.

Boedeker, P., & Kearns, N. T. (2019). Linear Discriminant Analysis for Prediction of Group Membership: A User-Friendly Primer. *In Advances in Methods and Practices in Psychological Science*, *2*(3), 250–263. https://doi.org/10.1177/2515245919849378

Breiman, L. (2001). Random Forests. *In Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Davies, E. R. (2018). Basic classification concepts. In E. R. Davies (Ed.), *: Computer Vision (Fifth Edition)* (pp. 365–398). Academic Press.

Edgar, T. W., & Manz, D. O. (2017). Machine Learning. In D. O. (senior C. S. S. P. N. Manz (Ed.), *: Research methods for cybersecurity: Syngress Media,u* (pp. 153–173). s.

Etemad, K., & Chellappa, R. (1997). Discriminant analysis for recognition of human face images. *In J. Opt*, *14*, 8. https://doi.org/10.1364/JOSAA.14.001724

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *In Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Friedman, J. H. (2002). Stochastic gradient boosting. *In Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gallardo, M. (2018). Identifying Vulnerability to Poverty: A Critical Survey. *Journal of Economic Surveys*, *32*(4), 1074–1105. https://doi.org/10.1111/joes.12216

Han, J., & Kamber, M. (2012). *Data mining. Concepts and techniques. 3rd ed*. Elsevier.

Izenman, A. J. (2013). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics.

Kambuya, P. (2020). Better Model Selection for Poverty Targeting through Machine Learning: A Case Study in Thailand. In *Undefined*. Available online at http://ethesisarchive.library.tu.ac.th/thesis/2017/TU_2017_5904040077_8907_8836.pdf

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & others. (n.d.). *(2017): LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. In Advances in Neural Information Processing Systems 30.

Lee, W., Lee, J., Lee, H., Jun, C.-H., Park, I.-S., & Kang, S.-H. (2014). Prediction of Hypertension Complications Risk Using Classification Techniques. *In Industrial Engineering and Management Systems*, *13*(4), 449–453. https://doi.org/10.7232/iems.2014.13.4.449

Martinez, A. M., & Kak, A. C. (2001). PCA versus LDA. *In IEEE Trans. Pattern Anal*, *23*(2), 228–233. https://doi.org/10.1109/34.908974

McLachlan, G. J. (2005). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons (Wiley series in probability and mathematical statistics. Applied probability and statistics). Available online at. http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10275312

Mohammed, S. S., & Al-Tuwaijari, J. M. (2021). Skin Disease Classification System Based on Machine Learning Technique: A Survey. *In IOP Conf. Ser.: Mater*, *1076*, 1. https://doi.org/10.1088/1757-899X/1076/1/012045

Nilashi, M., Ahmadi, N., Samad, S., Shahmoradi, L., Ahmadi, H., Ibrahim, O., Asadi, S., Abdullah, R., Abumalloh, R. A., & Yadegaridehkordi, E. (2020). Disease Diagnosis Using Machine Learning Techniques: A Review and Classification. *Journal of Soft Computing and Decision Support Systems*, *7*(1), 19–30.

Noori, B. (2021). Classification of Customer Reviews Using Machine Learning Algorithms. *In Applied Artificial Intelligence*, *35*(8), 567–588. https://doi.org/10.1080/08839514.2021.1922843

Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A comprehensive evaluation of ensemble learning for stock-market prediction. *In J Big Data*, *7*(1), 1–40. https://doi.org/10.1186/s40537-020-00299-5

Otok, B., & Seftiana, D. (2014). *The Classification of Poor Households in Jombang With Random Forest Classification And Regression Trees (RF-CART) Approach as the Solution In Achieving the 2015 Indonesian MDGs ' Targets*. Available online at. https://www.semanticscholar.org/paper/The-Classification-of-Poor-Households-in-Jombang-(-Otok-Seftiana/cc9acc7de20ceb9f4afd3f22573cf7323a97153d

Palaniappan, S., Mustapha, A., Foozy, M., Feresa, C., & Atan, R. (2017). Customer Profiling using Classification Approach for Bank Telemarketing. *In JOIV: Int*, *1*, 4–2. https://doi.org/10.30630/joiv.1.4-2.68

Pokhriyal, N., Zambrano, O., Linares, J., & Hernández, H. (2020). *Estimating and Forecasting Income Poverty and Inequality in Haiti Using Satellite Imagery and Mobile Phone Data*. Inter-American Development Bank. https://doi.org/10.18235/0002466

Ravikumar, S., & Saraf, P. (2020). Prediction of Stock Prices using Machine Learning (Regression, Classification) Algorithms. *In*, *2020*, 1–5.

Sammut, C., & Webb, G. I. (Eds.). (2010). *Encyclopedia of Machine Learning*. Springer US. https://doi.org/10.1007/978-0-387-30164-8

Schaefer, J., Lehne, M., Schepers, J., Prasser, F., & Thun, S. (2020). The use of machine learning in rare diseases: A scoping review. *In Orphanet J Rare Dis*, *15*, 1. https://doi.org/10.1186/s13023-020-01424-6

Thoplan, R. (2014). Random Forests for Poverty Classification. *In*, *1*(17), 252–259.

TNP2K. (2018a). Basis Data Terpadu. Available online at http://www.tnp2k.go.id/download/6713920180214-Buku%20Laporan%20BDT.pdf, checked on 6/17/2021.

TNP2K. (2018b). Laporan Evaluasi Pemanfaatan Basis-Data-Terpadu. Available online at http://tnp2k.go.id/download/29800Buku-Laporan-Evaluasi-Pemanfaatan-Basis-Data-Terpadu.pdf, checked on 6/17/2021.

Wang, P., Fan, E., & Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *In Pattern Recognition Letters*, *141*, 61–67. https://doi.org/10.1016/j.patrec.2020.07.042

Zixi, H. (2021). Poverty Prediction through Machine Learning. *Proc. - Int. Conf. E-Commer. Internet Technol., ECIT*, 314–324. Scopus. https://doi.org/10.1109/ECIT52743.2021.00073